

Predicting the Outcome of Patients With Subarachnoid Hemorrhage Using Machine Learning Techniques



Paula de Toledo, Pablo M. Rios, Agapito Ledezma, Araceli Sanchis, Jose F. Alen, and Alfonso Lagares

Abstract—Background: Outcome prediction for subarachnoid hemorrhage (SAH) helps guide care and compare global management strategies. Logistic regression models for outcome prediction may be cumbersome to apply in clinical practice. **Objective:** To use machine learning techniques to build a model of outcome prediction that makes the knowledge discovered from the data explicit and communicable to domain experts. **Material and methods:** A derivation cohort ($n = 441$) of nonselected SAH cases was analyzed using different classification algorithms to generate decision trees and decision rules. Algorithms used were C4.5, fast decision tree learner, partial decision trees, repeated incremental pruning to produce error reduction, nearest neighbor with generalization, and ripple down rule learner. Outcome was dichotomized in favorable [Glasgow outcome scale (GOS) = I–II] and poor (GOS = III–V). An independent cohort ($n = 193$) was used for validation. An exploratory questionnaire was given to potential users (specialist doctors) to gather their opinion on the classifier and its usability in clinical routine. **Results:** The best classifier was obtained with the C4.5 algorithm. It uses only two attributes [World Federation of Neurological Surgeons (WFNS) and Fisher’s scale] and leads to a simple decision tree. The accuracy of the classifier [area under the ROC curve (AUC) = 0.84; confidence interval (CI) = 0.80–0.88] is similar to that obtained by a logistic regression model (AUC = 0.86; CI = 0.83–0.89) derived from the same data and is considered better fit for clinical use.

Index Terms—Data mining, knowledge discovery in databases, machine learning, prognosis, subarachnoid hemorrhage.

I. INTRODUCTION

SPONTANEOUS subarachnoid hemorrhage (SAH) is a form of hemorrhagic stroke characterized by the presence of blood in the subarachnoid space, occupied by the arteries feeding the brain. The most common cause of SAH is the rupture of a cerebral aneurysm, an abnormal and fragile dilatation of a cerebral artery. Its annual incidence is 10–15 cases per 100 000 inhabitants and nearly 50% of patients suffering it will have a poor outcome. Brain damage related to this form of stroke is due to a decrease in cerebral blood perfusion leading to cerebral ischemia. Diagnosis is made with cranial computerized tomography (CT) scan that shows the extent of the bleeding. Cerebral angiography confirms the presence of an aneurysm in the ma-

This work was supported in part by the Spanish Ministries of Science under Grant TRA2007-67374-C02-02 and Health under Grant FIS PI 070152. The work of A. Lagares and J. F. Alen was supported by the Fundacion Mutua Madrileña.

P. de Toledo, P. M. Rios, A. Ledezma, and A. Sanchis are with the Control, Learning, and Systems Optimization Group, Universidad Carlos III de Madrid, Madrid 28040, Spain (e-mail: paula.detoledo@uc3m.es).

J. F. Alen and A. Lagares are with the Department of Neurosurgery, Hospital Doce de Octubre, Madrid 28041, Spain.

jority of the patients, although in nearly 20% of the cases the cause is unknown. The aneurysm, if found, should be treated as soon as possible as it has a natural tendency to rerupture (mortality over 50%). Treatment could be performed by endovascular means or surgically to exclude the aneurysm preserving normal circulation.

As in other acute neurological diseases, determining prognosis after SAH is crucial for giving adequate information to patient’s relatives, guide treatment options, detect subgroups of patients that could benefit from certain treatments, and compare treatments or global management strategies. Prognostic information coming just from surgically or endovascularly treated cases would not be applicable to all patients suffering this condition, as many patients die before being treated [1]. Therefore, any model valid for assessing prognosis at diagnosis in this disease should be obtained from a nonselected series of patients. Prognostic factors are mainly level of consciousness at admission, quantity of bleeding in the initial CT-scan, age, size of the aneurysm, and location [2], [3]. Different scales have been used to classify patients with SAH, with the World Federation of Neurological Surgeons (WFNS [4]) being the most frequently used. It divides patients in five grades according to the severity of consciousness disturbance. Its reliability and interobserver reproducibility are high, as it condenses the information coming from the Glasgow coma scale (GCS [5]), which is a universal scale for consciousness assessment. The amount of bleeding in the initial CT has been evaluated with different scales, some assessing the amount of cisternal blood in a qualitative way (Fisher’s scale [6]) and others using a semiquantitative algorithm [7]. The evaluation of the prognostic information given by these different scales has been done mainly with conventional statistics. Prognostic models have been built mainly for dichotomized six-month outcome using logistic regression analysis. Some scales have been built combining factors coming from these models, including age, Fisher’s scale, and WFNS [8], [9]. Their accuracy has been tested using the area under the receiver operating curve area under the ROC curve (AUC), achieving less than 90% accurate prognosis. The results are difficult to interpret in the clinical setting as they consist of different combinations of prognostic factors derived from several scales, combined by scores or coefficients derived from the regression equation. There is a need for simple, universal, interpretable, and reliable prognostic tools for SAH patients.

A. Data Mining in Prognosis

Predicting the future course and outcome of a disease process, as well as predicting potential disease onset on healthy patients,

is an active area of research in medicine. Prognostic models are primarily used to select appropriate treatments [10]–[13] and tests [14], [15] not only in individual patient management, but also in assisting comparative audit among hospitals by case-mix adjusted mortality predictions [16], guiding healthcare policy by generating global predictive scenarios, determining study eligibility of patients for new treatments, defining inclusion criteria for clinical trials to control for variation in prognosis, as well as in cost reimbursement programs.

Statistical techniques such as univariate and multivariate logistic regression analyses have been successfully applied to prediction in clinical medicine. A commonly used instrument is the use of a prognostic score derived from logistic regression to classify a patient into a future risk category. In the past ten years, decreasing costs of computer hardware and software technologies, availability of good quality high-volume computerized data, and advances in data mining algorithms, have led to the adoption of machine learning techniques approaches to a variety of practical problems in clinical medicine. A relevant summary of current research in the field, including techniques most widely used, can be found in a recent review paper by Belazzi and Zupan [17]. Other reviews that show the activity in progress are [11] and [18], where different techniques are presented and compared.

The question of whether artificial neural networks (ANNs) or other machine learning techniques can outperform statistical modeling techniques in prediction problems in clinical medicine does not have a simple answer. There are plenty of research works comparing techniques from the two domains [16], [19], [20], showing that there is no methodology outweighing the others in all possible scenarios, and that the tools need to be carefully selected depending on the problem faced and the significant quality criteria. In some cases, machine learning techniques have been shown to lead to similar results as logistic regression in accuracy, but outperform in calibration [13], [21]. Other authors highlight the ease of use and automation of techniques such as ANNs, while stating that logistic regression is still the gold standard [22]. Furthermore, statistical and machine learning techniques are not necessarily competing strategies, but can also be used together to perform a prediction task [16].

Most universally used predictive data mining methods, according to a poll conducted in 2006 among researchers in the field [23] are:

- 1) those based on decisions trees such as ID3 [24] and C4.5 [25];
- 2) those based on decision rules;
- 3) statistical methods, mainly logistic regression; and
- 4) ANNs, followed by support vector machines, naive Bayesian classifiers, Bayesian networks, and nearest neighbors.

Less used methods are ensemble methods (boosting, bagging) and genetic algorithms. In the field of prognosis in clinical medicine the results differ, as logistic regression is still the most widely applied, followed by ANNs [12]–[14], [20]. The use of decision trees [16], [19], [26] is growing in recent times. Other methods such as genetic algorithms are still scarce, but promising. [15], [21]. A growing trend is combining different

machine learning techniques to achieve improved results. [26]. When comparing different classifiers [17], [27], the key issues to address are:

- 1) predictive accuracy;
- 2) interpretability of the classification models by the domain expert;
- 3) handling of missing data and noise;
- 4) ability to work with different types of attributes (categorical, ordinal, continuous);
- 5) reduction of attributes needed to derive the conclusion;
- 6) computational cost for both induction and use of the classification models;
- 7) ability to explain the decisions reached when models are used in decision making; and
- 8) ability to perform well with unseen cases.

Interpretability of the results being the main selection criteria, besides accuracy, it is surprising that there is little research in the field. Harper [27] conducted a survey among the staff of set of NHS trusts in the south of U.K., comparing the comprehensibility and ease of use of models based on logistic regression, ANNs, and decision trees that concluded that the latter are the ones with a greater practical appeal. The interpretability of models obtained from logistic regression can be facilitated by the use of nomograms (Lubsen *et al.* [28]). Nomograms are a well-established visualization technique consisting of a graphic representation of the statistical model that incorporates several variables to predict a particular end point.

In the field of SAH, the classification and regression trees methodology (CART) has been compared to logistic regression analysis ($n = 885$) to predict the outcome of SAH patients [28]. Results obtained were similar and it was concluded that the single best predictor (level of consciousness) was itself as good as multivariate analysis. CART was also used in a similar condition [30], intracerebral hemorrhage ($n = 347$), to develop a classification tree that stratified the mortality in four risk levels and outperformed a multivariate logistic regression model in terms of accuracy (AUC 0.86 vs. 0.81).

B. Objectives

The aim of this paper is to use knowledge discovery and machine learning techniques to build a model for predicting the outcome of a patient with SAH, using only data gathered on hospital admission, which makes the knowledge discovered from the data explicit and communicable to domain experts, and which is usable in routine practice. To be usable, the model should use as few predictors as possible, be intuitive to interpret, and have a similar accuracy to techniques currently in use. The class attribute is the outcome six months after discharge, measured by means of the Glasgow outcome scale (GOS [31]), a five-point scale that is often dichotomized into “favorable outcome” and “poor outcome.” The main objective is to predict the dichotomized outcome, but models leading five and three (trichotomized) classes are also investigated.

TABLE I
CHARACTERISTICS OF COHORTS USED IN THIS STUDY

		Dataset1 Derivation cohort n=411, 1990-2001	Dataset2 Validation Cohort n=193, 2001-2007
Age		54,2	54,3
Female : Male		1,5:1	1:1*
Arterial hypertension		38,1%	39,4%
WFNS grade at admission	I	47,8%	47,2%
	II	22,4%	19,2%
	III	6,3%	3,6%
	IV	8,4%	11,4%
	V	14,9%	18,7%
Fisher Grade	I	8,8%	4,1%
	II	17,4%	21,8%
	III	37,4%	36,8%
	IV	36,2%	37,3%
Idiopathic SAH		26,7%	21,8%
Confirmed aneurysm		59,2%	70%*
Treatment	No	12%	22%*
	Treatment		
	Surgical	62%	31%*
	Endovascular	22%	43%*
Six month	Combined	3,8%	4%
	I-II	61,3%	59,1%
GOS	III-V	38,8%	40,9%

* Significant difference between cohorts ($P < 0,05$).

II. MATERIALS AND METHODS

A. Data Mining Methodology

The phases of the learning process presented in this paper correspond to those described by the Crisp-DM model [32], defined by the Cross-Industry Standard Process for Data Mining Interest Group. These phases are, with minor changes, common to most data mining methodologies: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The knowledge discovery process consists of a series of evolutionary cycles, covering one or more of those phases, repeating tasks such as data preparation, feature selection, selection of the data mining technique, generation of classifiers, and evaluation of the results.

B. Data Sources

We collected data retrospectively from two different data cohorts (Table I) holding information from all SAH patients admitted in a teaching hospital (Hospital Doce de Octubre) in Madrid, Spain. The first cohort (Dataset1) keeps information from 441 cases, from 1990 to 2001. The second (Dataset2) was created between 2001 and 2007 and has 192 cases, for which a smaller number of variables (a subset) were recorded. The strategy followed was to use the first dataset to select the attributes and train the classifier, and the second for external validation.

C. Business and Data Understanding, Data Preparation

Data gathered can be categorized as follows:

- 1) initial evaluation variables;
- 2) variables related to diagnostic cranial CT scan;
- 3) variables related to diagnostic angiography;
- 4) variables related to the type of treatment and level of consciousness before treatment; and
- 5) outcome variables, including complications (rebleeding, ischemia, vasospasm, etc.).

Outcome is measured, both at discharge and six months after. Only data available at the time of diagnosis (groups 1 and 2) are used for prediction, resulting in a total of 40 attributes.

The data were anonymized prior to its handing over to the research team, to comply with the Spanish national regulations on personal data. Informed consent had been obtained from all the patients before including their information in the registry.

D. Modeling

The open source tool Weka [33] was used in different phases of the knowledge discovery process. Weka is a collection of state-of-the-art data mining algorithms and data preprocessing methods for a wide range of tasks such as data preprocessing, attribute selection, clustering, and classification. Weka has been used in prior research both in the field of clinical data mining [34] and in bioinformatics [35].

1) *Attribute Selection*: Attribute selection is a key factor for success in the generation of the model. Different subset evaluators and search methods were combined. Subset evaluators used were classifier subset evaluator [33] (assesses the predictive ability of each attribute individually and the degree of redundancy among them, preferring sets of attributes that are highly correlated with the class but have low intercorrelation) and Wrapper [33] (employs cross validation to estimate the accuracy of the learning scheme for each attribute set). Search methods used were:

- 1) greedy stepwise [33] (greedy hill climbing without backtracking; stopping when adding or removing an attribute worsens the results of the evaluation, as compared to the previous iteration);
- 2) genetic search (using a simple genetic algorithm) [36];
- 3) exhaustive search [33] (exhaustive search in the attribute subset, starting from an empty set, and selecting the smallest subset); and
- 4) race search [33] (competitions among attribute subsets, evaluating them as a function of the error obtained in the cross validation).

2) *Classification Algorithms*: Among the different machine learning techniques available, decision trees and decision rules were preferred to neural networks for their interpretability. Decision trees, also called classification trees, are models made of nodes (leaves) and branches, where nodes represent classifications and branches correspond to conjunctions of features (values or value ranges) that lead a classification. The aim in decision tree learning is to use variables to partition the dataset into homogeneous groups with respect to the outcome variable (e.g., “favorable outcome,” “poor outcome”). The tree construction is achieved by recursively partitioning the dataset into subsets based on the value of a variable. In each iteration, the learning process looks for the variable leading to maximum homogeneity in the resulting subsets. Different measures of homogeneity can be used, resulting in different tree learning techniques. Decision rules are similar to decision trees, and can be derived from the former or produced directly, either from knowledge elicited from the experts or with machine learning techniques.

From the broad range of decision trees and decision rules algorithms available, the following were included in this study

according to their suitability to the problem domain: C4.5, fast decision tree learner (REPTree), partial decision trees (PART), repeated incremental pruning to produce error reduction (Ripper), nearest neighbor with generalization (NNge), ripple down rule learner (Ridor), and best-first decision tree learning (BFT). C4.5, REPTree, and BFT build decision trees whereas the rest are rule induction algorithms. C4.5 [25] is an improvement over ID3 [24], since it produces a decision tree using entropy to determine each tree node, but is not able to work either with incomplete data or with numerical attributes. C4.5 improves ID3, including the concept of gain ratio and admitting numerical attributes. REPTree [33] builds a decision tree by evaluating the predictor attributes against a quantitative target attribute, using variance reduction to derive balanced tree splits and minimize error corrections. PART [37] (obtaining rules from partial decision trees) is a rule induction algorithm. Such algorithms usually work in two phases: first, they generate classification rules, and then, these are optimized through an improvement process, usually with a high computational cost. PART algorithm does not perform such a global improvement, but uses the C4.5 algorithm to take the best leaf in each iteration and transform it into a rule. Ripper is a rule induction algorithm working in three phases as follows:

- 1) building (growing and pruning);
- 2) optimization; and
- 3) rule reduction.

It is an improved version of incremental reduced error pruning (IREP) [38]. NNge [39] is a nearest neighbor method of generating rules using nonnested generalized exemplars. Ridor [33] technique is characterized by the generation of a first default rule, using incremental reduced-error pruning to find exceptions to this rule with the smallest pondered error rate. In the second phase, the best exceptions are selected using the IREP algorithm. BFT [40] uses binary split for both nominal and numeric attributes, while for missing values, the method of “fractional” instances is used.

E. Evaluation

As it is possible to have a statistically but not yet clinically valid model and vice versa, evaluation must be conducted in two directions: laboratory evaluation of the performance of the model and clinical evaluation to determine whether the model is satisfactory for clinical purpose.

1) *Laboratory Evaluation:* Hit ratio and kappa statistics have been used to compare the different classifiers generated. Hit ratio is not a proper accuracy score, as it does not penalize models that are imprecise (for example, by exaggerating the probability of a dominant class). Kappa statistic [41] corrects the degree of agreement between the classifier’s predictions and reality by considering the proportion of predictions that might occur by chance, and is recommended [42] as the statistic of choice to compare classifiers. The receiver operating characteristics (ROC) curve [43] is another widely used tool. In the case of nonbinary classification problems, kappa statistic can be used as is, while ROC curves are more cumbersome to interpret.

Tenfold cross validation was used for internal validation. This well-known validation strategy is based on the partition

TABLE II
EXPERIMENTAL CONFIGURATION

Attribute selection methods: ClassifierSubsetEval, WrapperSubsetEval
Search methods: GreedyStepwise, GeneticSearch, ExhaustiveSearch, RaceSearch
Classification algorithms: C4.5, REPTree, PART, NNge, Ripper, Ridor, BTF
Class attribute:
Dichotomized: favorable outcome GOS=I-II / poor outcome GOS=III-V
Trichotomized: favorable GOS=I-II / severe disability GOS= III-IV / death GOS= V.
Training mode: Cross validation: 10 folds

of the original sample into ten subsamples, retaining one for testing and using the remaining nine as training data. The cross-validation process is then repeated ten times with each of the ten samples, averaging the results from the tenfolds to produce a single estimation. External validation of the best classifier was performed with an independent dataset (hold out strategy).

2) *Clinical Evaluation:* To assess the potential usefulness of the classifier in clinical routine, the results were presented to six neurosurgeons from five different hospitals in Spain. A questionnaire with 21 questions (five-point Likert scale) was prepared by the research team, covering issues related to the value of the model, interpretability, and potential use in clinical routine.

F. Deployment

It must be noted that the classifier developed is intended to be used as a support tool in a Web-based multicentric register of SAH cases. Therefore, it should be possible to implement it in a way that can be integrated with such technologies.

III. RESULTS

A. Modeling

The data mining process consisted of three iterative cycles, as described in the methodology. For each cycle, two different sets of experiments were performed: a first battery to select the more relevant attributes and a second battery to build the classifier itself. Table II shows the experimental configuration, including attribute selection, search, and classification algorithms used.

The experiments of the first two cycles resulted in:

- 1) the variables representing the amount of blood in the ten cisterns being substituted by a summary score;
- 2) continuous variables such as age being clustered; and
- 3) the outcome variable being clustered in two and three classes (Table II).

The experiments of the third cycle used 27 attributes. Different datasets were prepared according to the following:

- 1) nonaggregated attributes, so that the splitting values are set by the attribute selection algorithm;
- 2) attributes clustered as decided by the technical team;
- 3) attributes clustered as decided by the expert; and
- 4) only age clustered.

Attribute selection led to 38 datasets with attributes ranging from 1 to 23.

For the dichotomized problem, kappa values and hit ratio were very similar for C4.5, PART, REPTree, Ripper, and BTF

TABLE III
VALIDATION FOR THE BEST CLASSIFIER GENERATED BY EACH ALGORITHM

		C4.5	PART	BFT	REPTree	NNGe	Ripper	Ridor
DICHOT. PROBLEM	Hits	364	364	364	365	344	364	354
	%	82.5	82.5%	82.5%	82.8%	78.0%	82.3%	80.2%
	Kappa	0.630	0.630	0.630	0.633	0.531	0.630	0.580
	# attributes	2	3	2	3	5	2	7
TRICHOT PROBLEM	Hits	344	336	339	342	291	338	342
	%	78.0	76.2%	76.8%	77.5%	66.0%	76.6%	77.5%
	Kappa	0.556	0.517	0.530	0.545	0.343	0.519	0.551
	# attributes	2	4	3	3	3	3	7
PRECISION VALUES..								
		DICHOT. PROBLEM		TRICHOTOMIZED PROBLEM.				
		Favorabl e outcome	Poor outcome	Favorabl e outcome	Severe disabilit y	Death		
C4.5	TP Rate	0.87	0.754	0.878	0	0.781		
	FP Rate	0.246	0.13	0.269	0	0.168		
	Precision	0.848	0.787	0.837	0	0.677		
PART	TP Rate	0.87	0.754	0.87	0	0.737		
	FP Rate	0.246	0.13	0.304	0	0.174		
	Precision	0.848	0.787	0.819	0	0.656		
REPTree	TP Rate	0.878	0.749	0.881	0	0.759		
	FP Rate	0.251	0.122	0.292	0.005	0.155		
	Precision	0.846	0.795	0.826	0	0.689		

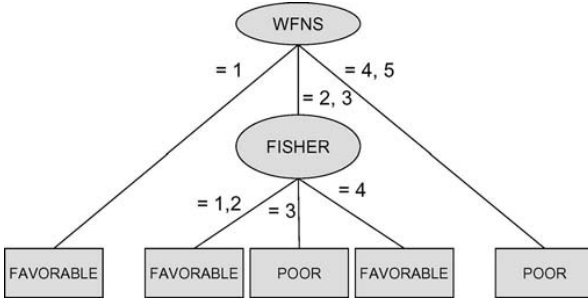


Fig. 1. Final classifier: C4.5 decision tree, dichotomized outcome.

models (Table III). Precision values were similar for the first three whereas NNGe, Ripper, and Ridor had slightly worse results. The best model was the one created by the C4.5 algorithm, which is shown in Fig. 1. It used only two attributes (WFNS and Fisher's) and had a lower complexity (six branches, five leaves) as compared to the others. The attributes selected for this classifier (WFNS and Fisher's scale) had been present in the experiments of all previous cycles, and their selection was consistent with the relevance assigned to them by the expert. The best model generated by the PART algorithm led exactly to good results, but the number of rules was higher (17) and more difficult to interpret according to the domain expert. The attributes selected (WFNS, Fisher's, and number of previous hemorrhages) were also consistent from the clinical point of view.

Regarding the results for the trichotomized problem, the percentage of correctly classified instances was only slightly lower than those obtained for the dichotomized scale (Table III); however, a more careful insight considering the kappa statistic, confusion matrix, and precision values for each class, showed that the intermediate class (severe disability) was very defectively classified. As there were very few cases (34) of this class, this had a small effect in the overall hit ratio but a great impact on

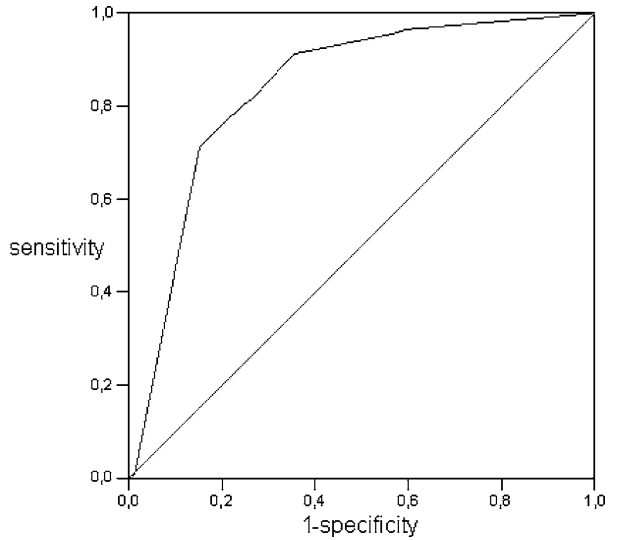


Fig. 2. ROC curve for the final classifier.

the utility of the classifier. Attributes used by the best model were WFNS and the score summarizing amount of blood in the cisterns. To handle the imbalance of the dataset, a further experiment was performed replicating the instances in the intermediate class (three times) to increase their overall weight in the classifier learning process. Results were slightly better (79% hit ratio, 0.670 kappa), but the number of instances in the intermediate class, which were correctly classified, is still only four (out of 34). A further experiment was performed attending to a request by the expert: generate a trichotomized tree with the same attributes used by the dichotomized (WFNS and Fisher's). Results were 1% hit ratio, 0.476 kappa.

The model chosen is therefore the one created by the C4.5 algorithm for the dichotomized problem, a tree with six branches and five leaves, shown in Fig. 1. The quality values for this classifier are AUC = 0.841 [0.80–0.88; confidence interval (CI) 95%], hit ratio = 83%, and kappa = 0.625.

B. External Validation

As the attributes selected by the model were present in Dataset2 (Table I), it was possible to perform an external validation of the selected classifier with this independent test set. The results obtained were AUC = 0.837 (0.78–0.89; 95% CI), 78% hit ratio, 0.73 sensitivity (for the “poor outcome” class), 0.81 specificity, and 0.55 kappa. The ROC curve is shown in Fig. 2.

External validation was performed as well using a random subset of the two datasets both for training and testing. The classifier generated was the same (Fig. 1) and the results were only slightly better: 80% hit ratio, 0.73 sensitivity, 0.86 specificity, and 0.60 kappa. This indicates that it is possible to generate a classifier from cases available at a certain point of time that preserves its predicting ability for future patients.

TABLE IV
RESULTS OF MULTIVARIATE LOGISTIC REGRESSION ANALYSIS ($B = -0.46$;
HOSMER-LEMESHOW GOODNESS OF FIT; $\chi^2 = 6.03$; DOF = 8; $p = 0.65$)

	β	Relative risk of poor outcome (Exp β)	CI	p value
Age	0.58	1.06	1.06-1.08	<0.01
WFNS grade ^a	II	1.23	1.8-6.6	<0.01
	III	1.77	2.2-15.5	<0.01
	IV	2.89	6.7-48.9	<0.01
	V	4.35	28-217	<0.01
Fisher grade ^b	II	0.15	0.2-6.7	=0.89
	III	1.77	1.3-27.4	<0.05
	IV	1.04	0.6-13.4	=0.12

CI: confidence interval.

^aCompared with WFNS grade 1.

^bCompared with Fisher's grade 1.

C. Statistical Model

The results of the multivariate logistic regression analysis using factors recorded at admission for dichotomized outcomes are shown in Table IV. A backward stepwise strategy was used to build the model. The attributes selected are: WFNS, Fisher's grade, and age. The AUC for the model in the derivation cohort is 0.86 (0.83–0.89, 95% CI). Coefficients from logistic regression models are difficult to interpret and different strategies have been used in order to calculate individual probabilities, such as converting these models into a score or using nomograms. Such a strategy was used and a nomogram was plotted from the logistic regression model results (Fig. 3).

1) *Clinical Evaluation*: Six neurosurgeons responded to the exploratory questionnaire. None of them had used a model based on machine learning techniques before. They considered the model simple to understand (4.0 in a five-point Likert scale), and sound from the clinical point of view (4.0). As compared to logistic regression models, they found it easier to interpret (3.7) and reported similar trust on the methodology (3.0). The fact that the model used only two variables was considered an advantage (3.8). All respondents agreed that the classifier could be used in clinical routine (4.3) and that integrating it both into the hospital information systems (3.7) and in the multicenter registry (3.8) would be a plus.

D. Deployment: Integration With the Multicenter Register

The Web-based multicentric register was modified to add a “show prognosis” button, which displays the graphical presentation of the decision tree highlighting the branch that led to the classification. In order to cope with future changes in the classifier, the system reads the graph from a standard graph representation based on DOT templates. This is the format used by the open source Weka library [33], whose graphical implementation package we modified to work as a Java applet. This applet is able to accept any decision tree and represent it. As a drawback of this implementation, we note that it requires the installation of the Java runtime environment in the user's computer. The classifier itself is coded in Visual Basic and the time needed to offer a prediction is negligible (mean time to load the page <1 s).

IV. DISCUSSION

The C4.5 algorithm leads the best model using automatic learning techniques. PART produced similar results, but the clinical interpretation of the tree was more obscure. The accuracy of the classifier, expressed in terms of AUC is 0.84 (0.80–0.88; 95% CI), in the range of the results [AUC = 0.86 (0.83–0.89, 95% CI)] obtained with the logistic regression model [8] (range 0.83–0.86 for the different scales studied). As compared to the logistic regression model, the decision tree uses one factor less (both use WFNS and Fisher's grade, logistic model uses age as well).

WFNS is known to be the best single predictor of outcome [8], [28]. Although age has been repeatedly found to be a determinant prognostic factor in SAH, when using conventional statistics [2], [44], it must be pointed out that the results obtained with the C4.5 algorithm when the age attribute is added to the filtered learning data are worse than when it is ignored.

It could have been expected that the classifier would show a linear progress from “favorable outcome” to “poor outcome.” Conversely, it can be noticed (Fig. 1) that for Fisher's grade 3, results are worse than those for grade 4. This lack of linearity in the Fisher's scale has been found by other researchers [2], [9] and resides in the very definition used when assessing Fisher's grade of subarachnoid bleeding. In this scale, when a thick clot is found in the subarachnoid space, a grade 3 is always assigned by definition. Therefore, the proportion of patients with vasospasm and cerebral ischemia, and therefore, poor outcome is higher for Fisher's grade 3 than for grade 4 [6].

The use of a nonselected series of patients is the main value of this paper as compared to Germanson [29], who used data from patients selected for a randomized trial where the presence of an aneurysm confirmed by angiography was needed for inclusion. Many patients with diagnosed SAH die before angiography (nearly 10% in our series) and also many patients with SAH do not harbor an aneurysm (more than 20% in our series). Therefore, data and prognostic information from selected cases are not applicable to all SAH patients at diagnosis. In Germanson's work, patients are stratified in three levels of risk for unfavorable outcome, although there is no assessment of the accuracy of their prediction in terms that allow for comparison with our results.

The diagnostic capability of the decision tree equals that of the logistic regression model, while the tree brings about some advantages that are as follows:

- 1) a decision tree is more intuitive and simpler to interpret than a nomogram;
- 2) it contains a reduced number of rules;
- 3) uses one factor less (age); and
- 4) is easier to generate.

The experts interviewed agreed on the fact that the decision tree is easier to interpret than the nomogram and gave to the two models the same diagnostic capability. When presented with the question “the predictions achieved using logistic regression are more trustworthy than those obtained using machine learning,” only two of the respondents “moderately agreed,” whereas the other four either disagreed or were neutral. According to our

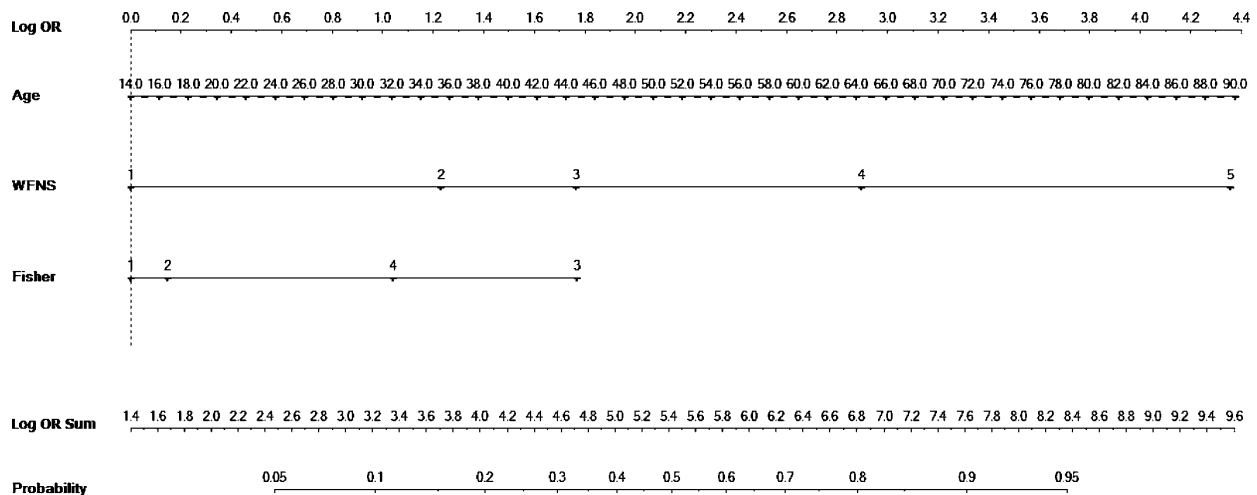


Fig. 3. Nomogram summarizing information derived from the logistic regression model (using the derivation cohort), showing probabilities of poor outcome.

exploratory survey, potential applicability of this model in clinical practice is high. Integration of the classifier into a multicenter registry or into the information systems used in clinical practice was very positively valued. We expected a lack of confidence on machine learning techniques from the experts, but this was not the case. It must be noted that the survey is very limited ($n = 6$), and therefore, these conclusions are only indicative, but as none of these experts had used machine learning techniques before, a favorable bias in this group is not foreseen.

The results obtained for the trichotomized problem (favorable outcome, severe disability, death) were useless from the clinical point of view. The number of cases in the dataset that belong to the intermediate class is small (34 cases), and therefore, the learning process does not succeed with this subset. Interobserver variability in data collection is another very well-known source of error for these models. Current work with statistical methods is mainly for a dichotomized outcome variable, which is usable from a clinical point of view, although the greater challenge for the future stands in achieving a good prediction for intermediate cases. Future work will target the improvement of the results in this area, working with higher number of cases (likely coming from the multicentric registry) and with other machine learning algorithms (for example, combined classifiers such as boosting and bagging or genetic algorithms). Another field with potential for improvement is the prediction of complications (such as rebleeding, hydrocephalus, or vasospasm), the prediction of outcome depending on treatment, and for different patient subpopulations.

The results are limited by the size of the training set (441 instances). However, SAH is a relatively rare condition and building larger databases is not always possible. A further limitation of the results is that they have been derived from patients from a single hospital, and therefore, its applicability outside this organization is unknown. Before integrating the classifier into the multicenter registry, the model should be tested and improved, if necessary, with data gathered from all hospitals involved, in order to increase its generalization ability.

The open source toolkit Weka has proved to be a very useful instrument supporting the data mining process. Furthermore, the tree coding format used by Weka has been used to integrate the classifier into the multicenter registry. The use of a standard language to represent the classifier that can be interpreted by the information system used in routine and modified without having to change the information system itself is an interesting way to facilitate the adoption of decision support tools in clinical practice. An alternative format is the Predictive Data Mining Markup Language [45], a vendor-independent open standard that defines an XML-based markup language for the encoding of many predictive data mining models, including decision trees and logistic regression. A further step would be to configure the prediction tool as a Web service offered to information systems subscribing to it. The model could be incrementally learning from the new cases introduced in the multicenter registry and offer updated decision support information online to electronic healthcare record systems from different healthcare providers. Researchers in the field of artificial intelligence in medicine agree that the impact in clinical practice of prognosis tools is maximized when these are made accessible through computer-based systems that are integrated into the clinician's workflow [16], [46]. The integration of the prediction model into the multicenter registry is a step, yet only investigative, toward this goal.

REFERENCES

- [1] H. Saveland, J. Hillman, L. Brandt, G. Edner, K. Jakobson, and G. Algers, "Overall outcome in aneurysmal subarachnoid hemorrhage. A prospective study from neurosurgical units in Sweden during a 1-year period," *J. Neurosurg.*, vol. 76, pp. 729–734, 1992.
- [2] A. Lagares, P. A. Gomez, R. D. Lobato, J. F. Alen, R. Alday, and J. Campollo, "Prognostic factors on hospital admission after spontaneous subarachnoid hemorrhage," *Acta Neurochir. (Wien)*, vol. 143, pp. 665–672, 2001.
- [3] H. Saveland and L. Brand, "Which are the major determinants for outcome in aneurysmal subarachnoid hemorrhage? A prospective total management study from a strictly unselected series," *Acta Neurol. Scand.*, vol. 90, pp. 245–250, 1994.

- [4] C. G. Drake, W. E. Hunt, K. Sano, N. Kassell, G. Teasdale, B. Pertuiset, and J. C. Devilliers, "Report of the World Federation of Neurological Surgeons committee on a universal subarachnoid hemorrhage grading scale," *J. Neurosurg.*, vol. 68, pp. 985–986, 1988.
- [5] G. Teasdale and B. Jennett, "Assessment of coma and impaired consciousness. A practical scale," *Lancet*, vol. 2, no. 7872, pp. 81–84, Jul. 1974.
- [6] C. M. Fisher, J. P. Kistler, and J. M. Davis, "Relation of cerebral vasospasm to subarachnoid hemorrhage visualized by computed tomographic scanning," *Neurosurgery*, vol. 6, pp. 1–9, 1980.
- [7] A. Hijdra, P. J. A. M. Brouwers, M. Vermeulen, and J. van Gijn, "Grading the amount of blood on computed tomograms after subarachnoid hemorrhage," *Stroke*, vol. 21, pp. 1156–1161, 1990.
- [8] A. Lagares, P. A. Gomez, J. F. Alen, R. D. Lobato, J. J. Rivas, R. Alday, J. Campollo, and A. G. de la Camara, "A comparison of different grading scales for predicting outcome after subarachnoid haemorrhage," *Acta Neurochir. (Wien)*, vol. 147, no. 1, pp. 5–16, Jan. 2005.
- [9] C. S. Ogilvy and B. S. Carter, "A proposed comprehensive grading system to predict outcome for surgical management of intracranial aneurysms," *Neurosurgery*, vol. 42, pp. 959–970, 1998.
- [10] H. Seker, M. O. Odetayo, D. Petrovic, and R. N. Naguib, "A fuzzy logic based-method for prognostic decision making in breast and prostate cancers," *IEEE Trans. Inf. Technol. Biomed.*, vol. 7, no. 2, pp. 114–122, Jun. 2003.
- [11] L. Ohno-Machado, F. S. Resnic, and M. E. Matheny, "Prognosis in critical care," *Annu. Rev. Biomed. Eng.*, vol. 8, pp. 567–599, 2006.
- [12] G. F. Cooper, V. Abraham, C. F. Aliferis, J. M. Aronis, B. G. Buchanan, R. Caruana, M. J. Fine, J. E. Janosky, G. Livingston, T. Mitchell, S. Monti, and P. Spirtes, "Predicting dire outcomes of patients with community acquired pneumonia," *J. Biomed. Inf.*, vol. 38, no. 5, pp. 347–366, Oct. 2005.
- [13] Y. C. Li, L. Liu, W. T. Chiu, and W. S. Jian, "Neural network modeling for surgical decisions on traumatic brain injury patients," *Int. J. Med. Inf.*, vol. 57, no. 1, pp. 1–9, Jan. 2000.
- [14] B. A. Mobley, E. Schechter, W. E. Moore, P. A. McKee, and J. E. Eichner, "Neural network predictions of significant coronary artery stenosis in men," *Artif. Intell. Med.*, vol. 34, no. 2, pp. 151–161, Jun. 2005.
- [15] M. Buscema, E. Grossi, M. Intraligi, N. Garbagna, A. Andriulli, and M. Breda, "An optimized experimental protocol based on neuro-evolutionary algorithms application to the classification of dyspeptic patients and to the prediction of the effectiveness of their treatment," *Artif. Intell. Med.*, vol. 34, no. 3, pp. 279–305, Jul. 2005.
- [16] A. Abu-Hanna and N. de Keizer, "Integrating classification trees with local logistic regression in intensive care prognosis," *Artif. Intell. Med.*, vol. 29, no. 1/2, pp. 5–23, Sep./Oct. 2003.
- [17] R. Bellazzi and B. Zupan, "Predictive data mining in clinical medicine: Current issues and guidelines," *Int. J. Med. Inf.*, vol. 77, no. 2, pp. 81–97, Feb. 2008.
- [18] P. J. Lucas and A. Abu-Hanna, "Prognostic methods in medicine," *Artif. Intell. Med.*, vol. 15, no. 2, pp. 105–119, Feb. 1999.
- [19] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: A comparison of three data mining methods," *Artif. Intell. Med.*, vol. 34, no. 2, pp. 113–127, Jun. 2005.
- [20] G. Clermont, D. C. Angus, S. M. DiRusso, M. Griffin, and W. T. Linde-Zwirble, "Predicting hospital mortality for patients in the intensive care unit: A comparison of artificial neural networks with logistic regression models," *Crit. Care Med.*, vol. 29, no. 2, pp. 291–296, Feb. 2001.
- [21] F. Jaimes, J. Farbiarz, D. Alvarez, and C. Martinez, "Comparison between logistic regression and neural networks to predict death in patients with suspected sepsis in the emergency room," *Crit. Care*, vol. 9, no. 2, pp. R150–R156, Apr. 2005.
- [22] R. Linder, I. R. Konig, C. Weimar, H. C. Diener, S. J. Poppl, and A. Ziegler, "Two models for outcome prediction—A comparison of logistic regression and neural networks," *Methods Inf. Med.*, vol. 45, no. 5, pp. 536–540, 2006.
- [23] (2006). KDNuggets Data Mining Methods Poll [Online]. Available: http://www.kdnuggets.com/polls/2006/data_mining_methods.htm
- [24] R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.
- [25] R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1993.
- [26] Z. H. Zhou and Y. Jiang, "Medical diagnosis with C4.5 rule preceded by artificial neural network ensemble," *IEEE Trans. Inf. Technol. Biomed.*, vol. 7, no. 1, pp. 37–42, Mar. 2003.
- [27] P. R. Harper, "A review and comparison of classification algorithms for medical decision making," *Health Policy*, vol. 71, no. 3, pp. 315–331, Mar. 2005.
- [28] J. Lubsen, J. Pool, and E. van der Does, "A practical device for the application of a diagnostic or prognostic function," *Methods Inf. Med.*, vol. 17, no. 2, pp. 127–129, Apr. 1978.
- [29] T. P. Germanson, G. Lanzino, G. L. Kongable, J. C. Torner, and N. F. Kassell, "Risk classification after aneurysmal subarachnoid hemorrhage," *Surg. Neurol.*, vol. 49, no. 2, pp. 155–163, Feb. 1998.
- [30] O. Takahashi, E. F. Cook, T. Nakamura, J. Saito, F. Ikawa, and T. Fukui, "Risk stratification for in-hospital mortality in spontaneous intracerebral haemorrhage: A classification and regression tree analysis," *QJM*, vol. 99, no. 11, pp. 743–750, Nov. 2006.
- [31] B. Jennett and M. Bond, "Assessment of outcome after severe brain damage," *Lancet*, vol. 1, no. 7905, pp. 480–484, Mar. 1975.
- [32] C. Shearer, "The CRISP-DM model: The new blueprint for data mining," *J. Data Warehousing*, vol. 5, no. 4, pp. 13–22, 2000.
- [33] I. H. Witten and F. Eibe, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Francisco, CA: Morgan Kaufmann, 2005.
- [34] M. H. Ou, G. A. West, M. Lazarescu, and C. Clay, "Dynamic knowledge validation and verification for CBR teledermatology system," *Artif. Intell. Med.*, vol. 39, no. 1, pp. 79–96, Jan. 2007.
- [35] J. E. Gewehr, M. Szugat, and R. Zimmer, "BioWeka—Extending the Weka framework for bioinformatics," *Bioinformatics*, vol. 23, no. 5, pp. 651–653, Mar. 2007.
- [36] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, 1st ed. Reading, MA: Addison-Wesley, 1989.
- [37] E. Frank and I. H. Witten, "Generating accurate rule sets without global optimization," in *Proc. 15th Int. Conf. Mach. Learn.* San Francisco, CA: Morgan Kaufmann, 1998, pp. 144–151.
- [38] W. W. Cohen, "Fast effective rule induction," in *Proc. 12th Int. Conf. Mach. Learn. (ML 1995)*, pp. 115–123.
- [39] B. Martin, "Instance-based learning: Nearest neighbor with generalization," Master's thesis, Univ. Waikato, Hamilton, New Zealand, 1995.
- [40] S. Haijian, "Best-first decision tree learning," Ph.D. dissertation, Univ. Waikato, Hamilton, New Zealand, 2007.
- [41] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.
- [42] A. Ben-David, "What's wrong with hit ratio?" *IEEE Intell. Syst.*, vol. 21, no. 6, pp. 68–70, Nov./Dec. 2006.
- [43] J. A. Hanley, "Receiver operating characteristic (ROC) methodology: The state of the art," *Crit. Rev. Diagn. Imag.*, vol. 29, pp. 307–335, 1989.
- [44] G. Lanzino, N. F. Kassell, T. P. Germanson, G. L. Kongable, L. L. Truskowski, J. C. Torner, and J. A. Jane, "Age and outcome after aneurysmal subarachnoid hemorrhage: Why do older patients fare worse," *J. Neurosurg.*, vol. 85, no. 3, pp. 410–418, Sep. 1996.
- [45] Data Mining Group. (2006). The Predictive Model Markup Language (PMML) [Online]. Available: www.dmg.org
- [46] M. Stefanelli, "The socio-organizational age of artificial intelligence in medicine," *Artif. Intell. Med.*, vol. 23, no. 1, pp. 25–47, Aug. 2001.